学校编码：**10384** 分类号______密级______

学　　号：**20051300791** UDC _____

厦门大学

硕　士　学　位　论　文

半参数变系数分位数回归模型及其两阶段估计：以波士顿房价应用为例

Semiparametric Functional Coefficient Quantile Regression and Its Two-step Estimation Procedure：An Application to Boston Housing Prices Data

翁　云　妹

指导教师姓名：蔡宗武　教授
　　　　　　　方颖助理教授
专 业 名 称：计量经济学

论文提交日期：2008 年 4 月

论文答辩时间：2008 年　月

学位授予日期：2008 年　月

答辩委员会主席：__________

评　　阅　　人：__________

2008 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　　　　　日期：　　年　月　日

导师签名：　　　　　　日期：　　年　月　日

摘 要

半参数模型、变系数模型和分位数回归是当今回归分析中研究的热点问题。

20世纪70年代，Koenker和Bassett提出了著名的线性分位数回归模型，有效地解决了数据分布中存在厚尾或异常值的难题。但在实际应用中，线性分位数回归无法捕捉潜在的较为复杂的分位数结构，如一些变量之间存在着高度的非线性相关或相互影响的关系。为此，学者们提出了非参数分位数回归模型，以观测出变量间复杂的非线性关系，然而该模型却无法观测出变量间的线性关系。

本文在半参数模型、变系数和分位数回归的基础上，提出了一种新的半参数变系数分位数回归模型。半参数变系数分位数回归模型，同时含有线性和非线性两个部分。该模型既可以观测到变量间复杂的线性或非线性等关系，还可以捕捉到每个变量的具体变动情况。由于分位数回归的非线性性质，传统的近似最小二乘估计法（Profile Least Square）并不能直接运用到半参数变系数分位数回归模型中。因此，在半参数变系数分位数回归模型的估计上，本文提出了一个新的两阶段估计方法。对于模型中线性部分的系数估计，我们首先将模型中线性部分的变量系数当作非参数形式，利用非参数估计法估计出变量的系数函数，然后对系数函数求均值。对于模型中非线性部分的系数估计，我们用理论带宽选择法选出一系列可能的带宽并以非参数AIC 为标准从中选出一个最优的带宽，然后选择一般的高斯核函数和局部多项式线性估计法对变量的系数函数做出估计。

最后，本文以具有典型性和权威性的波士顿房价数据作为半参数变系数分位数回归模型及其两阶段估计应用的一个例子，展现该模型在现实经济问题中良好的解释能力。

关键词：分位数回归；变系数；两阶段估计

Abstract

At present, semiparametric model, functional coefficient and quantile regression are key issues in the field of regression analysis.

Koenker and Bassett proposed a famous linear quantile regresion which solves heavy tail problems effectively in 1970s. But in application, the model can not catch a more complicated quantile structure, such as nolinear or interactive relationships among variables. Then researchers propose nonparametric quantile regressions to get more information. However, nonparametic quantile regression can not catch the linear relationship among variables.

In this paper, we propose a new semiparametric quantile regression. based on reveiews of semiparametric model, functional coefficient and quantile regression. Semiparametric quantile regression contains linear and nonlinear part simultaneously, it not only observes complicated realtionship among variables, but also discovers detailed movement for each variable. We also propose a new two-step estimation procedure to estimate the semiparametric quantile regression. Because traditional profile least square is invalid for nonliearity of quantile regression. To estimate the constant parameters in the model, first, the constant coefficients are regarded as functional coefficients and then we apply a nonparametric estimation procedure. The final estimators of those parameters are obtained by the average method. To estimate the functional coefficents optimally, we simple use the partial residuals, then apply a nonparametic estimation procedure.

Finally, we take a famous boston housing prices data as an example to show the power of semiparametric quantile model and its two-step estimation in application.

Key Words: Quantile Regression; Functional Coefficient; Two-step Estimation Procedure

目 录

Contents

第一章 绪论

第一节 研究背景和目的

在过去的 30 年间，由 Koenker 和 Bassett (1978) [1][2]提出的分位数回归模型（Quantile Regression Model），已成功地并广泛地应用于各个领域，如金融、经济、医药和生物行业。尤其是，非参数和半参数分位数回归模型倍受关注。因为在理论和实证上，非参数和半参数分位数模型都比线性分位数模型具有更强的灵活性，例如 Chaudhuri，Doksum 和 Samarov (1997) [3]，Yu 和 Lu (2004) [4]，Cai 和 Xu (2005) [5]。

具体而言，令$\{V_t, Y_t\}_{t=-\infty}^{\infty}$为一组平稳序列，$F(y\,|\,v)$是在$V_t = v$下$Y_t$的条件分布，其中$V_t$是一组变量向量，包含有可能的外生变量（Exogenous Variable）和滞后变量（Lagged Variable）。对于任意的$0<\tau<1$，分位数回归函数$q_\tau(v)$为

$$q_\tau(v) = \inf\{y \in \Re : F(y\,|\,v) \ge \tau\} \quad \text{或} \quad q_\tau(v) = \arg\min_{a\in\Re} E\{\rho_\tau(Y_t - a)\,|\,X_t = v\}$$

其中，$\rho_\tau(y) = y[\tau - I(\{y<0\})]$称为损失函数，$I(g)$为指数函数。同时，$q_\tau(v)$称为条件分位数（Conditional Quantile）或分位数回归（Regression Quantile）。

分位数回归包括线性分位数回归和非线性分位数回归。在许多实际应用中，线性分位数回归无法捕捉潜在的较为复杂的分位数结构。例如，一些变量之间存在高度的非线性相关关系或者变量之间存在着相互影响的关系。非线性分位数回归主要指的是非参数分位数回归。在高维度分析中，非参数分位数回归存在着“维度祸根”（Curse of Dimensionality）问题，使其在实际应用中很难解释现实经济现象。一些可减少维度的模型构建法，如变系数分位数回归模型（Functional Coefficient Quantile Regression），来解决非参数分位数回归中的“维度祸根”问题，如 Yu 和 Lu (2004) [4]，Cai 和 Xu (2005) [5]。但是，非参数分位数回归模型却无法捕捉到变量间的线性相关关系。

本文研究的主要目的在于提出了一种新的半参数变系数分位数回归模型，在该模型中，一些外生的和滞后的连续或离散变量的系数可以是线性的，也可以是非线性的。这样，该模型就可以观测到变量间复杂的线性或非线性等关系，同时还可以灵活地捕捉到每个变量的具体变动情况。由于分位数回归的非线性性质，传统的近似二乘

估计法（Profile Least Square）不能够直接运用到半参数变系数分位数回归模型中。因此，本文提出了一种新的两阶段估计方法对半参数变系数分位数回归模型做出估计。

半参数变系数分位数回归模型的最大优点在于，可以考虑任何有关变量线性性和非线性的信息，因为该模型本身就包含有线性和非线性两个部分。此外，半参数变系数分位数回归模型，还综合了半参数、变系数和分位数回归三种模型的优点，如需要较少的样本数据，估计时线性部分的收敛速度较快，还可以解决数据中的非正态分布和异常值等问题，使得该模型在实际经济应用中具有很好的灵活性和广泛的适用空间。

第二节 研究对象和方法

本文试图在回顾半参数、变系数和分位数回归模型及其估计的基础上，提出一种新的模型和估计方法——半参数变系数分位数回归模型及其两阶段估计。为了展现该模型对于实际经济问题的良好解释能力，本文选取了具有代表性的波士顿房价作为该模型应用的一个例子。

本文提出了两阶段估计对半参数变系数分位数回归模型做出估计，先估计出模型中线性部分的系数，再估计出模型中非线性部分的系数函数。对于模型中线性部分的系数估计，我们首先将其当作是非线性的并对其进行估计，即非参数变系数分位数模型的估计，然后对估计出的系数函数求均值。对于模型中非线性部分的系数估计，我们采用理论带宽选择法选出一系列可能的带宽值后，再用 Cai 和 Xu(2005)[5]提出的偏差校正后的非参数 AIC 标准选出一个最优的带宽，然后选择一般的高斯核函数和局部多项式线性估计法对参数做出估计。

本文将半参数、变系数模型和分位数回归这些在回归分析中较为灵活的方法结合到一个模型中并对其进行估计，是本文的创新之处。

第三节 本文结构安排

本文提出了一类新的模型——半参数变系数分位数回归模型，还提出了两阶段估计对该新模型做出估计。

第二章回顾了半参数、变系数模型和分位数回归及其估计的发展历程以及研究现

况。

首先大体上介绍了非参数回归和半参数回归的发展历程，非参数估计中的各种核函数和带宽的选取方法，估计的统计性质及其研究现状。在非参数回归理论的基础上，介绍了变系数回归理论的具体模型及其估计方法。然后介绍分位数回归的相关理论及其研究现况。最后，在变系数回归和分位数回归的基础上，进一步讨论了变系数分位数回归模型和局部多项式线性估计法。

第三章提出了半参数变系数分位数回归模型的一般形式和两阶段估计的基本思想。

第四章介绍了波士顿房价数据的来源、结构和属性，并回顾了有关波士顿房价的研究文献。我们给出了一些波士顿房价数据的描述性统计表和图形，还进行了一些变量的有偏性检验。

第五章以波士顿房价数据为例，阐释了如何应用两阶段估计对半参数变系数分位数回归模型做出估计。

首先对模型中的变量进行筛选，其次，我们对波士顿构建了变系数分位数回归模型，选择高斯核函数和非参数 AIC 标准选出一个最优的带宽，利用局部多项式线性估计法对模型进行估计。然后，基于波士顿房价的属性，我们构建了半参数变系数分位数回归模型，采用两阶段估计对模型进行重新估计。

第六章给出了本文的最终结论，证实了半参数变系数分位数回归模型及其两阶段估计对实际经济问题具有很好的解释能力。

第二章 理论阐述和文献综述

第一节 非参数回归和半参数回归

一、理论阐述

非参数回归模型包括完全非参数回归模型(简称非参数回归模型)和半参数回归模型两类。参数回归最基本的假定是被解释变量的条件期望是依赖于解释变量和未知参数的已知函数，通常这些函数关系是线性的。即

$$Y = X\beta + \varepsilon$$

其中 X 为随机向量，β 为未知参数向量，ε 为随机误差向量。

但是，在现实中，变量之间的关系未必是线性关系或可线性化的非线性关系，变量之间的参数非线性关系即使存在也很难确定，在大多数情况下可供选择的函数是很多的，因而就难以避免模型设定误差的出现。而非参数回归并不假定固定的回归形式，它假定变量之间的函数关系是未知的，但首先要先对这个回归函数进行估计，因而从这个意义上来说非参数回归模型是较参数回归模型更符合实际的回归模型。

非参数回归是 20 世纪 30 年代中后期开始形成，并逐步发展起来的。它是与参数回归相比较而存在，不依赖于总体分布及其参数，亦即不受变量分布约束的模型。20 世纪 70 年代以来，非参数回归日渐兴起。非参数回归模型的特点是回归函数的形式可以任意，对解释变量 X 和响应变量 Y 的分布限制很少，因而有较大的适应性。之后，非参数回归模型的各种估计方法也就不断完善起来。自非参数回归估计的权函数（Weighted Function）估计法提出来后，该方法引起了广泛的重视。近几十年来，权函数方法如近邻估计（Neighbour Estimation）、核估计（Kernel Estimation）、局部多项式估计（Local Polynomial Estimation）等方法不断发展完善，非参数回归的理论和应用取得了较大的进展。

非参数回归也有其局限性，如需要大量的样本数据，收敛速度较慢和估计不稳定等问题。并且在非参数回归模型中，各个解释变量对响应变量的作用的差别往往被忽略，当实际问题对此未提供任何信息时，这是不可避免的；但若有根据(经验或历史资料等)认为某些解释变量对响应变量 Y 的影响较显著时，使用非参数回归模型则没有充分利用已知的信息，会明显降低模型的解释能力。而非参数和参数回归模型的结

合——半参数回归模型，可以弥补非参数回归模型的这个缺陷。

半参数回归模型的研究始于八十年代，半参数回归模型一出现，就引起了应用工作者和理论工作者的广泛重视。半参数回归模型理论研究主要集中在大样本的性质上，研究热点主要集中在回归参数估计方法的建立，回归参数估计算法的建立，收敛性问题的讨论和回归模型诊断等问题。至今，对于半参数回归中未知函数的处理方法无论在理论上还是在实际应用中都还不太成熟，即在理论上还未形成完整的体系，在估计方法上比较零乱，而且计算过程较复杂。

目前，常用的半参数回归有以下五种模型：

$$
\begin{aligned}
&Y = X\beta + g(T) + \varepsilon \\
&Y = \beta + g(T) + \varepsilon \\
&Y = X\beta + g(X) + \varepsilon \\
&Y = f(X,\beta) + g(T) + \varepsilon \\
&Y_i = Y_{i-1}\beta + g(Y_{i-2},\cdots,Y_{i-p}) + \varepsilon
\end{aligned}
$$

其中，$(X,T)\in R^p \times R^1$ 为随机向量，T 的支撑集为有限闭集，β 为 $p\times 1$ 维的未知参数向量，$g(\cdot)$ 是定义在一有限闭集上的未知函数，ε 为随机误差向量。且 $E(\varepsilon)=0$，$E(\varepsilon^2)=\sigma^2$，$\varepsilon$ 与 (X,T) 相互独立。

半参数回归模型具有许多优点，如其需要较少的样本数据，线性部分的收敛速度较快且本身具有经济含义。

二、非参数估计：核函数和带宽选择

在非参数估计中，由于密度函数是未知的，故要估计出密度函数的形式。密度函数的估计涉及到两个重要的概念：核函数和带宽（Bandwidth）选择。

设 $K(x)$ 为一个概率密度函数，$h>0$ 为给定的常数，$f(x)$ 为总体的密度函数，则 $\hat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n}k(\frac{x-X_i}{h})$ 称为 $f(x)$ 的核估计。其中，函数 $K(x)$ 称为核（Kernel），常数 h 称为带宽（Bandwidth）。

核函数通常满足对称性及 $\int k(x)dx=1$，并且在原点上只有单峰的密度函数。核密度估计的实质是对样本点施加不同的权数，用加权来代替通常的记数，核函数即为核

权函数。而核估计利用了数据点x_i到x的距离$x-x_i$来决定x_i在估计点x的密度函数估计，离x越近的点所赋予的加权越大；反之，则加权越小。h称带宽（满足$h\to 0, nh\to\infty$），h越大，估计出的密度函数则越平滑，但偏差也越大。在实际的应用中，常用的核函数$K(x)$，见表 1。

表 1　常用的核函数

核 函数的名称	核函数（$-\infty<u<\infty$）
均匀核	$K(u)=\begin{cases}0.5,-1\le u\le 1\\ 0, \quad other\end{cases}$
三角核	$K(u)=(1-\|u\|)$
Epanechikov 核	$K(u)=0.75(1-u^2)_+$
四次方核	$K(u)=\frac{15}{16}((1-\|u\|^2)_+)$
六次方核	$K(u)=\frac{70}{81}((1-\|u\|^3)_+)$
高斯核	$K(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$
余弦核	$K(u)=\begin{cases}\frac{1}{2}\cos(u),\|u\|\le\frac{\pi}{2}\\ 0, \quad other\end{cases}$

通常，核函数的大样本估计具有以下性质:

（1）核估计的渐近无偏性

设$x_1,x_2\cdots,x_n$是一组具有密度函数f的总体样本。那么核估计密度函数$\hat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n}k(\frac{x-X_i}{h})$是$f$的渐近无偏估计，即$\lim_{n\to\infty}E\,\hat{f}_n(x)=f(x)$。

（2）核估计的均方相和性